DOCUMENT RESUME

ED 236 154 TM 820 341

AUTHOR	Hale, Gordon A.; And Others
TITLE	Effects of Item Disclosure on TOEFL Performance.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-80-34; TOEFL-RR-8
PUB DATE	Dec 80
NOTE	52p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Achievement Gains; *College Entrance Examinations; Educational Legislation; *Foreign Students; Scores; *Test Coaching; Test Construction; *Test Wiseness
IDENTIFIERS	*Test Disclosure; *Test of English as a Foreign Language

ABSTRACT

To ascertain how the Test of English as a Foreign Language (TOEFL) would be affected if candidates had access to some of the items before administration of a test containing those items, a number of specially constructed TOEFL forms were made available to 945 foreign students in intensive English language programs. The students were later administered a special TOEFL consisting of items from those forms and a TOEFL consisting of undisclosed items. A significant disclosure effect was determined when performance was contrasted on these two tests. Scores for the test containing disclosed items were greater than those for the test containing undisclosed items, indicating that students studied the disclosed forms and recalled specific items for them. Student questionnaire data suggested that the time spent studying the disclosed material was relatively unaffected by the number of forms given the students. Additional analyses indicated disclosure effects for items in forms that were not discussed in class as well as for items that were discussed. A disclosure effect was observed for each of three main language groups and for every item type, which attests to the robustness of this effect. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED236154

TOEFL Research Reports

REPORT 8
DECEMBER 1980

EFFECTS OF ITEM DISCLOSURE ON TOEFL PERFORMANCE

Gordon A. Hale
Paul J. Angelis
Lawrence A. Thibodeau

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Educational Testing Service

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of over thirty organizations, public and private, that were concerned with testing the English proficiency of non-native speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program and in 1973 a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; Graduate Record Examinations Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the Graduate Record Examinations Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

Effects of Item Disclosure on TOEFL Performance

Gordon A. Hale

Paul J. Angelis

Lawrence A. Thibodeau

Educational Testing Service
Princeton, N.J.

RR 80-34

Copyright © 1980 by Educational Testing Service. All rights reserved.

Unauthorized reproduction in whole or in part is prohibited.

ABSTRACT

New legislation requires that standardized tests be disclosed after they have been administered. According to procedures followed to date, TOEFL forms previously used in International or Special Center administrations have been provided to institutions for use in the Institutional testing program. In light of the new legislation, alternative procedures need to be explored. One alternative is to combine items from previously used forms, possibly including ones that have been disclosed, into new forms for use in the Institutional program. Before initiating such a plan it is necessary to know how TOEFL performance would be affected if candidates had access to some of the items before administration of a test containing those items.

The present study addressed this question through an experimental approach. A number of specially constructed TOEFL forms, here called "disclosed forms," were made available to foreign students in intensive English language programs. Later the students were administered a special TOEFL consisting of items from those forms and a TOEFL consisting of undisclosed items. Effects of item disclosure were determined by contrasting performance on these two tests.

The students were given copies of the disclosed forms, and their language instructors devoted class time to discussing a portion of the items in those forms. The study was thus expected to show the effect that might be produced if, as might happen in reality, students were tutored on disclosed items in test-preparation courses. Separate analysis of test performance for items that had been disclosed but not discussed in class would show the extent to which students benefit from disclosed items when left on their own initiative to study them.

An additional variable in the study was the size of the disclosed item pool. It was hypothesized that, if students must cover a large number of test forms in order to be exposed to all the items that will appear on a later test, they will be less likely to benefit from disclosure than if they need only cover a small number of forms. To test this hypothesis, items to appear on the test were spread through six disclosed forms for students in eight institutions and through twelve forms for students in eight other institutions.

A significant disclosure effect was observed: Scores for the test containing disclosed items were greater than those for the test containing undisclosed items, indicating that the students studied the disclosed forms and recalled specific items from them. The effect of disclosure was greater for students who received six disclosed forms than for those who received twelve forms, confirming the hypothesis presented above. The effect for students receiving the larger number of forms was significant, nevertheless. Student questionnaire data suggested that the time spent studying the disclosed material was relatively unaffected by the number of forms given the students.

Additional analyses indicated disclosure effects for items in forms that were not discussed in class as well as for items that were discussed. Thus, the effect of disclosure was not due solely to the class experience, as the students apparently studied the disclosed forms on their own initiative. A disclosure effect was observed for each of three main language groups and for every item type, which attests to the robustness of this effect.

ACKNOWLEDGEMENTS

This study was designed in collaboration with Donald Alderman and Donald Powers, with advice from Russell Webster and Joan Borum. Their substantial contribution to the study is gratefully acknowledged.

Sincere appreciation is also expressed to Francean Meredith and Nancy Thomas for their efforts in selecting test items and compiling the many TOEFL forms used in this study.

The authors also wish to extend their thanks to:

Bruce Kaplan and Ingeborg Stiebritz for programming the data analyses; Sherrill Lord, Leta Davis and John Dunn for assistance in data analysis; Linda Dellaria and Ronald Zollars for aid in compiling the test forms; Jessie Cryer and Mary Beth Brookshaw for secretarial assistance; and Henry Braun, Brent Bridgeman and Marilyn Hicks for general advice and for a critical reading of this manuscript.

The authors are especially grateful to the persons at the participating institutions who made this study possible. Listed below are those who coordinated the study at each institution. Sincere thanks are expressed to these persons and to all the instructors who devoted their time to the study.

American University, English Language Institute
Mrs. Mary Ann Hood

Arizona, University of, Center for English as a Second Language
Ms. Phyllis Lim

Colorado State University, Intensive English Program
Dr. James Bachmann

Delaware, University of, English Language Institute
Ms. Patricia Dyer and Mr. Timothy Collins

Florida, University of, English Language Institute; in conjunction with Santa Fe Community College
Dr. Jayne Harder and Mr. Michael Pyle

George Washington University, English for International Students
Dr. George Bozzini

Georgetown University, Division of English as a Foreign Language
Dr. William Norris and Mr. Marvin Kierstead

Georgia State University, English as a Second Language Program
Ms. Rebecca Bodnar

Houston, University of (Downtown College), Intensive English Institute
Mr. Nicholas Franks

Houston, University of (Language and Culture Center), Intensive Language Program
Dr. Joyce Valdes and Mr. Michael Bettler

Illinois, University of, Intensive English Institute
Dr. Rebecca Dixon

Iowa State University, Intensive English and Orientation Program
Dr. William Flick and Ms. Jacqueline Saban

Louisville, University of, Intensive English as a Second Language Program
Mr. Patrick Kameen

Miami, University of, Intensive English Program
Mr. John Rogers and Mr. John Stevenson

Michigan State University, English Language Center
Dr. Paul Munsell and Mrs. Doris Scarlett

North Texas State University, Intensive English Language Institute
Dr. John Crow and Mrs. Nancy Strickland

Ohio University, Ohio Program of Intensive English
Dr. Robert Dakin

Temple University, Intensive English Language Program
Mrs. Rebecca Lemaitre

Texas A & M University, English Language Institute
Mrs. Jean Erb

Tulsa, University of, English Institute for International Students
Dr. Delores Bedingfield

TABLE OF CONTENTS

	Page
INTRODUCTION	1
METHOD	5
Subjects	5
Materials	5
Procedure	7
Experimental Design	10
RESULTS	13
Scoring of Tests	13
Computation of Test Reliability	13
Determination of Sample for Data Analysis	13
Analyses of Total Test Scores	14
Analysis for Items Covered in Class	20
Analysis for Items Not Covered in Class	21
Analysis by Language Group	22
Analysis by Test Section	22
Analysis of Student Questionnaire Data	24
Instructor Questionnaire Data--Overview	25
DISCUSSION	27
Overall Disclosure Effect	27
Effects of Class Coverage and Independent Study	29
Variation in Effect Due to Size of Disclosed Item Pool	30
Other Results	31
Implications for the TOEFL Program	31
REFERENCES	33
APPENDICES	
A. Analysis of Data from Control Institutions	35
B. Computation of Score for Items Covered in Class	37
C. Analysis of Instructor Questionnaire Data	39
D. Calculation of Scaled-Score Estimate of Disclosure Effect	43

LIST OF TABLES

Table		Page
1.	Experimental Design	11
2.	Schedule of Events for Each Institution	12
3.	Mean Posttest Scores for Each Institution	15
4.	Adjusted Mean Disclosed-Undisclosed Difference Scores for Each Condition	18
5.	Mean Posttest Scores for Each Control Institution	19

INTRODUCTION

The impetus for this study came from legislation recently instituted in the state of New York (effective January 1, 1980) and pending in a number of other states. The principal requirements of the New York law are that copies of standardized tests must be filed with the state after they have been administered and that persons who have taken those tests be able to request copies of the test they took and their own answer sheets. In the fall of 1979, when the present study was undertaken, many of the details regarding these requirements had not yet been determined. Specific regulations were being prepared within New York state which would address issues such as which tests fell under the legislation, which administrations would be covered, and to which candidates test materials would have to be supplied. Even without these details, however, the overall "disclosure" situation created problems of long-range if not short-range planning for most testing programs at ETS.

For the TOEFL program there was less concern for what the impact of such legislation would be on its primary operations--International and Special Center administrations. Although a total of twelve such administrations are offered each year (six of each type), a completely new form is prepared for each. Therefore, items from one administration do not appear on future forms of the test. Furthermore, the initial calendar of test dates prepared to comply with the legislation called for a maximum of five forms to be disclosed in the first year. What was of greater concern for TOEFL was the Institutional program. Under this program forms previously used at International or Special Center administrations are provided on request to institutions largely for use in making placement decisions. Although each of these forms has been used in one previous International or Special Center administration, materials are returned to ETS and a distribution plan is followed which is meant to minimize the possibility of previous exposure to test items prior to an Institutional administration. Under the requirements of the new legislation, however, many forms previously eligible for use in the Institutional program can no longer be used since they will have been disclosed and thus made a matter of public record.

A possible alternative is not to reuse complete forms but to combine items from previously used forms into new ones for administration in the Institutional program, perhaps even items from disclosed forms. Before initiating such a plan it would be necessary to know what effect there would be on TOEFL performance if candidates did have access to some of the items before the test administration. The present study was conducted to address this question. An experimental approach was used in an effort to simulate the way test forms might be made available in reality. A principal assumption was that, in practice, disclosed TOEFL forms will become subject to general distribution through study groups, test preparation courses, and other channels. Therefore, it was reasoned that the effects of disclosure might best be studied in the

context of a situation in which disclosed test forms are highly accessible. Hence, conditions were experimentally created in which disclosed test forms were maximally available to students before they took a TOEFL. The principal question to be addressed was whether, and to what extent, the students' TOEFL scores would be increased by having these disclosed forms available for study.

Specifically, a number of specially constructed TOEFL forms were given to foreign students enrolled in intensive English programs at sixteen universities throughout the country. The students were told that a two-part TOEFL testing session would be held later and that this special test would contain items from the disclosed forms. A few weeks later the students were administered a TOEFL consisting of items drawn from the disclosed forms and a second TOEFL consisting of undisclosed items (with order of presentation reversed at half the institutions). Comparison of performance on these two tests would provide the basis for inferring effects of disclosure; to conclude that TOEFL performance is affected by the prior availability of items, scores on the test containing disclosed items would need to be significantly higher than those on the test with no disclosed items.

It seems reasonable that, as a result of the disclosure law, disclosed items not only will become readily available but may well become subject to direct tutoring. Thus, at least some items from disclosed forms will likely become incorporated into the material used in test-preparation classes--whether formal coaching courses or supplementary classes in English language programs. Use of these items could result either (a) from instructors' intention to expose students to the specific items that will appear on a TOEFL, or (b) simply from the use of publicly available test forms as teaching aids. In either case, if disclosed forms become included among test-preparation materials, then students may receive class tutoring on items they will later encounter in an Institutional TOEFL (if items from disclosed forms are indeed reused). Therefore, it is necessary to determine the extent to which TOEFL performance is affected by in-class instruction on the items that appear on the test. Toward this end, a tutoring situation was created in this study by having the students' English language instructors devote approximately five hours to discussing portions of the disclosed forms. By assessing the effects of such instruction on later test performance, inferences could be drawn about the effects that might be expected if students are tutored on disclosed test items.

It is important not only to assess the effects of direct instruction with disclosed items but also to determine the extent to which students benefit from disclosure of items when they are left on their own initiative to study them. In the present study, the majority of the disclosed items were contained in forms that were not designated for class discussion. Separate analyses for these items alone would therefore allow inferences as to the effect of disclosure for items that students must study on their own initiative.

An additional factor believed to influence the magnitude of the disclosure effect is the size of the disclosed item pool. It was hypothesized that, if students must cover a large number of test forms in order to be exposed to all items that will appear on a later test, they will be less likely to benefit from disclosure than if they need only cover a small number of test forms. Two factors underlie this hypothesis. First, it is believed that students would be less likely to cover all disclosed forms with a larger item pool than with a smaller one because it would take a greater amount of time to do so. Second, even if the students were to cover all items, a larger item pool would place a greater demand on memory than would a smaller item pool. To test the hypothesis, students in eight institutions were given six disclosed forms, containing a total of 900 items, while students in the other eight institutions were given twelve disclosed forms, containing a total of 1800 items. Items that would appear on the special TOEFL were spread throughout the disclosed forms in each case. A significant difference in disclosure effect between these two groups would support the hypothesis.

METHOD

Subjects

The sample consisted of foreign students enrolled in intensive English language programs at twenty institutions of higher education in the United States. These noncredit programs are provided for students whose level of proficiency is generally below that required to obtain entrance into a regular academic program. It was left to the participating institutions to decide which groups of students in their programs to include in the study, with the request that a reasonably wide range of English proficiency be represented. In most cases, only those groups with the lowest levels of proficiency were excluded from the study, since these students' command of the English language is so poor that they would not likely be among those taking the TOEFL at this stage of their language training.

A total of 945 students participated in the study, and the final sample on which the principal analyses were based consisted of 668 students, the number on hand for all of the tests to be described below. A TOEFL administered as a pretest at the outset of the study (see "Materials" below) yielded mean scaled scores per institution ranging from 388 to 494 with an average of 437 across the twenty institutions (based on students in the final sample). The average age of the students in the final sample was 25 years, with a range of 17 to 66 years (a range of 19 to 37 years, excluding the top 5% and bottom 5%). Sixty-eight percent of the students in the final sample were males and 32% were females.

The predominant native languages represented were Spanish (37% of the final sample), Arabic (19%), Farsi (7%), Japanese (6%), Chinese (5%), Vietnamese (3%), Thai (3%), French (3%), Korean (3%), and Portuguese (2%). The predominant native countries of Spanish-speaking students were Venezuela (15% of final sample), Mexico (5%), Columbia (5%), and Bolivia (3%); the predominant native countries of Arabic speakers were Jordan (6% of final sample), Saudi Arabia (4%), and Lebanon (4%); the native countries corresponding to the other languages listed above were, respectively, Iran (7%), Japan (6%), China (5%), Vietnam (3%), Thailand (3%), Ivory Coast (3%), Korea (3%), and Brazil (1%).

Materials

General format of the TOEFL. The TOEFL consists of 150 total items in three sections: I Listening Comprehension; II Structure and Written Expression; and III Reading Comprehension and Vocabulary. Section I is divided into three subsections, containing 20, 15, and 15 items respectively. In each subsection, the examinee hears a number of tape-recorded speech segments to which he or she is to respond by selecting answers from among those printed in the test booklet. The speech segments in the first

subsection are short statements; those in the second are short dialogues; and those in the third are longer dialogues or brief monologues. For Sections II and III all questions and answers are printed in the test booklet. Section II consists of two subsections containing 15 and 25 items, respectively. In the first subsection each question requires the examinee to identify the phrase that best completes a sentence, while in the second subsection each question calls for identifying the underlined word or phrase that is ungrammatical. Section III consists of two 30-item subsections. In each item of the first subsection the examinee must choose the synonym for the word or expression underlined in a sentence. In the second subsection, the examinee must read some pieces of prose material and answer several questions about each.

Pretest. To establish the students' initial levels of proficiency, an operational form of the TOEFL currently employed for institutional administrations was used as a pretest. Performance on this test was not intended to be compared directly with performance on the tests given later; rather, the pretest was included to allow statistical control for initial proficiency levels.

Posttests. Two TOEFL forms, labeled here Form A and Form B, were specially constructed for use as posttests--i.e., the special TOEFLs to be administered at the end of the study. They were constructed from retired operational items taken from forms of the five-section TOEFL, which was in use until 1976 (only item types that are still in use were considered). The selected items in certain sections differed in minor respects from items in current operational TOEFLs; for example, about half of the reading passages used in this study were nonacademic in nature, whereas all passages in current TOEFLs contain material encountered in an academic context. Nevertheless, all items used here were identical in format to those of a current operational TOEFL, constructed according to the descriptions provided above. Forms A and B of the test were designed to be similar to each other with respect to essential characteristics such as item difficulty, length of reading passages, and overall statistical design.

Disclosed forms. The disclosed forms made available for use during the study also consisted of 150 items each, with the same numbers of items in the seven subsections as in an operational TOEFL. They were comprised of retired operational items from the five-section TOEFL. The physical layout of each form was the same as that of an operational TOEFL except that (a) the cover page read: "Test of English as a Foreign Language--QUESTIONS" rather than "TOEFL--Test of English as a Foreign Language," and (b) the correct answer to each question was marked with an asterisk. For each disclosed form, the material for the Listening Comprehension section was recorded on tape.

There were two major conditions in the study, which differed in the number of disclosed forms distributed to the students. In the 6-form condition, students received six disclosed forms containing a

total of 900 items. In the 12-form condition, students received twelve disclosed forms containing a total of 1800 items. Interspersed throughout the disclosed forms in each case were a number of items that would appear on the disclosed posttest. In the 6-form condition there were 25 such items in each form; in the 12-form condition there were between 10 and 14 such items per form. The forms in the 6-form condition were constructed by printing on 11" x 17" sheets, folding and stapling in the center; forms in the 12-form condition were printed on 8" x 11" sheets and stapled in the left-hand margin. The forms used in these two conditions were thus identical in size and were identical in appearance except for placement of the staple.

Separate sets of disclosed forms were constructed, one containing items that would appear in posttest Form A but not Form B, and the other containing items that would appear in posttest Form B but not Form A.[1] Specifically, the disclosed forms were constructed as follows. For the 12-form condition, two different sets of twelve disclosed forms were devised. One set was constructed from a bank of 1650 items plus the 150 items that would appear in posttest Form A. The other set was constructed from 1650 items plus the 150 items that would appear in posttest Form B. For the 6-form condition two different sets of six disclosed forms were devised according to a similar procedure except that, in this case, 750 rather than 1650 items were combined with the 150 items to appear in a posttest.

Procedure

Pretest. All students were administered a TOEFL as a pretest to establish their initial levels of proficiency. This test was administered in the same manner as an Institutional TOEFL with one difference. The students were told that, if they did not wish to have their test scores reported to either themselves or their institution, they could so indicate on their answer sheets. This instruction, given with each posttest as well, was necessary to minimize any perception of coercion. Few students elected this option--13 for the pretest and 12 for the posttests. Those who did were excluded from the sample, since many students in this group may have addressed the test in a less serious manner than they would if trying to maximize their scores.

Distribution of disclosed forms. After the pretest was administered the disclosed forms were made available to students in the principal

---

[1] It was necessary to construct the materials in such a way that for some institutions posttest Form A would be the test containing items that had been seen and Form B would be the test containing unseen items, while for other institutions the reverse would be true. (See Experimental Design at the end of this section.)

conditions of the study--usually within a week. (Students in the control institutions received no disclosed forms, as explained below under "Experimental Design.") The students were told that after a specified number of weeks there would be a special TOEFL testing session in two parts and that in this testing session they would encounter items from the disclosed forms. Depending on the condition to which an institution was assigned, the students received either six or twelve disclosed forms. All students were given copies of every printed form to keep in their possession during the study.

The audio portions of the disclosed forms (for the Listening Comprehension section) were recorded on tapes, and copies of these tapes were made available to the students. Each institution established a plan whereby students could check out tapes from the language laboratory, library, or classroom and keep them for a period of time ranging from two days to a week. Enough copies were made available so that, at any time, each student could have one tape. Thus, for example, if there were 60 students participating in the study at a given institution, and if that institution were among those given six disclosed test forms, then ten copies of each of six tapes were made available for distribution, or a total of 60 tapes.

The students were free to listen to these tapes on their own recorders or in a central facility in which tape recorders were made available. In addition, tapes were on file in the central facility for use only at that location, to further ensure that all students would have an opportunity to listen to the tapes. At every institution, 24 such tapes were on file--two copies of each of the twelve tapes for the 12-form condition, or four copies of each of the six tapes for the 6-form condition.

The disclosed test forms and tapes were made available to the students for a disclosure period of a few weeks. (The schedule of events at each institution is presented in Table 2 on page 12.) The students were allowed to keep the disclosed forms through the time the second posttest was presented; they could not reasonably be expected to relinquish the materials before this time. Hence, those students who took the disclosed posttest second had additional time to benefit from these materials. The experimental design (described at the end of this section) compensated for this effect by randomly assigning the institutions to different testing orders.

Class coverage of disclosed items. During the period in which the disclosed forms were available, certain groups of items from these forms were to be discussed in class. At some institutions this discussion was conducted in regular classes, and at other institutions it was conducted in special sessions held outside of the regular classes. Approximately five hours was to be devoted to this discussion.

Instructors were asked to discuss the items in Section I of one disclosed form and the items in Sections II and III of another disclosed form. In effect, then, the equivalent of one full test form was to be discussed in class; it was necessary to involve two forms rather than one, however, to achieve a proper representation of item types. (Any given form contained either items from Subsection I-3 that would appear on the posttest or items from Subsection III-2 that would appear on the posttest, but not both, since items in these two subsections occurred in clusters.)

A principal objective of the class coverage was to ensure that the students would be exposed to all designated items. Hence, the instructor was to have the students read (or listen to, for Section I) all designated questions and note the correct answers. Beyond this guideline, the manner in which the items were to be covered was left to the instructor's discretion. For example, the instructor could (a) have the students read (listen to) groups of items and then initiate discussion, or (b) have the students read (listen to) and discuss each item in turn. The instructor was also free to field questions from the students, offer spontaneous observations, or both. Further, the instructor could discuss selected aspects of the items--e g., similarities among items of a given type, principles of grammar represented in the items, or other aspects. In general, the objective was to simulate the circumstances that would arise if students were participating in test-preparation or tutoring groups, with the assumption that each such group would deal with the materials in its own way. Instructors were asked to respond to a brief questionnaire at the end of the study to obtain information about the nature and extent of class coverage.

Posttests. After the disclosed forms had been available for a few weeks (or after a delay of a few weeks for the control groups) two TOEFL posttests were administered. In most cases the tests were administered one week apart (see schedule in Table 2 below.) One of the posttests, entitled the "disclosed posttest," consisted entirely of items selected from the disclosed forms. The other, the "undisclosed posttest," consisted entirely of undisclosed items. Comparison of performance on those two posttests provided a basis for inferring effects of item disclosure.

The posttests were administered in the same manner as an Institutional TOEFL except that some of the biographical information that had been requested when the pretest was administered, such as native country and native language, was not requested again. Also, the students were informed that this was a special TOEFL testing session that would yield only raw scores (i.e., number correct per test section, and total number correct) rather than scaled scores. The students were told that these scores would not be used for any purposes for which Institutional TOEFL scores are normally used. This disclaimer was necessary to dispel any assumptions that performance on this experimental TOEFL would have a

bearing on academic advancement. Nevertheless, it was expected that the students would be highly motivated to score well on these tests. Most program directors reported that whenever a TOEFL is given in their program, whether an Institutional TOEFL or a practice test, the students are strongly motivated to do as well as possible.

Experimental Design

Principal conditions. There were sixteen institutions in the main conditions of the study. Table 1 on page 11 shows the design of the study. At eight institutions the students were given six disclosed test forms, and at eight other institutions the students were given twelve disclosed forms. Four institutions in each group of eight received the disclosed posttest followed by the undisclosed posttest, and the other four received the posttests in the reverse order. Thus, there were four experimental subgroups in all, defined by the combination of two factors, number of disclosed forms and test order. The sixteen institutions were randomly assigned to these four experimental subgroups.

As noted above, two test forms were constructed for use as posttests, labeled here Form A and Form B. For half the institutions, Form A served as the disclosed posttest and Form B the undisclosed posttest, and for the other half, the reverse was true. As shown in Table 1, form was intentionally confounded with order of presentation, so that for institutions receiving the disclosed posttest first, Form A was the disclosed posttest, and for institutions receiving the disclosed posttest second, Form B was the disclosed posttest. (Hence, the posttests were presented in order Form A-Form B for all institutions; the design assumes that the basic results would be approximately the same if the posttests were presented in order Form B-Form A rather than in this order.)

Control institutions. Four control institutions were included to provide independent data on order effects and on the comparability of the two posttests. Students in these institutions were first given the pretest, then after a delay of a few weeks they were given the two posttests without disclosure of any items. The posttests were presented in the order Form A-Form B at two institutions and the reverse order at the other two institutions. It should be noted that the control institutions were not included for purposes of direct comparison with the institutions in the principal conditions but to contribute data to aid in interpretation of the principal effects.

Test scheduling. Table 2 on page 12 presents the schedule of events for each institution. It can be seen that the time between distribution of disclosed materials and first posttest was just under two weeks for one institution and ranged from three-and-a-half to six weeks for the others. The time between posttests was usually one week. (Entries of zero in the last column indicate that the two posttests were administered on the same day, one in the morning and one in the afternoon.)

Table 1

Experimental Design

6-form disclosure condition		12-form disclosure condition	
Disclosed posttest administered first (Form A)	Disclosed posttest administered second (Form B)	Disclosed posttest administered first (Form A)	Disclosed posttest administered second (Form B)
Institution A	Institution E	Institution I	Institution M
Institution B	Institution F	Institution J	Institution N
Institution C	Institution G	Institution K	Institution O
Institution D	Institution H	Institution L	Institution P

Control institutions

Posttests administered in order Form A then Form B	Posttests administered in order Form B then Form A
Institution Q	Institution S
Institution R	Institution T

Table 2

Schedule of Events for Each Institution

Institution	Days from pretest to first posttest	Days from distribution of disclosed forms to first posttest	Days from first posttest to second posttest
A	13	12	1
B	49	34	7
C	42	38	7
D	28	28	7
E	40	39	7
F	28	27	7
G	42	40	7
H	41	33	7
I	42	25	7
J	35	28	7
K	42	42	7
L	41	28	7
M	45	28	7
N	77	37	7
O	28	26	0
P	37	31	2
Q	35	--	0
R	25	--	9
S	20	--	6
T	14	--	2

RESULTS

Scoring of Tests

The principal data from the study were scores on the posttests. Each test was scored by taking the percent correct for each of the three test sections and computing the average. The result was a percent correct for the total test which gives equal weight to the three sections. This method of computing the total score parallels the method used in computing scores for an operational TOEFL, wherein scores for each section are weighted equally in computing an overall scaled score. Scaled scores could not be computed for the posttests since scaling data were not available for these specially constructed tests. For the pretest, although scaled scores were available a percent-correct score was computed for use in the data analyses to ensure use of a common metric for all tests.[2]

Computation of Test Reliability

Analyses were performed to establish the reliability of the tests constructed here for use as posttests. Since Form A was administered first and Form B second at eighteen of the institutions (i.e., all but control institutions S and T), the present analyses were based only on data from those eighteen institutions. To establish the reliability of a test form, only the data for students administered that form as the undisclosed posttest (plus students in control institutions Q and R) were examined. The numbers of students involved in these computations were 371 for Form A and 414 for Form B. Alpha-coefficient reliabilities were computed for each section of the test as well as for the entire test. For Form A the reliabilities were: Section I, .87; Section II, .77; Section 3, .84; whole test, .93. For Form B the corresponding figures were .87, .70, .81, and .91.

Determination of Sample for Data Analysis

The total sample consisted of those students who took the pretest, were assigned to a class for distribution and discussion of disclosed materials (except in the control institutions), and took at least one posttest. This excluded students who took the pretest but were not

---

[2]Although there is not a direct linear relation between percent-correct pretest scores and scaled pretest scores, these types of scores are highly related. (Institutional means for each of these two scores were calculated and the correlation between the two scores, with institutional mean as the unit of analysis, was .998.) Thus, the use of percent-correct scores for the pretest should yield results quite similar to those that would be produced if scaled scores were used.

participating in the study, which was a large group at institutions in which the pretest substituted for a regular administration of the TOEFL. Also excluded were a few students who indicated a desire not to have their scores reported (see "Procedure" above). The final sample was defined as those students in the total sample who took all tests. This latter group was about 70% of the total sample, since it was not possible to ensure that all students would take all tests.

It was planned that inferences would be drawn from this final sample, provided that the students in the final sample were representative of the total sample with respect to their level of English proficiency. To address this question, pretest scores were computed (a) for students in the final sample, and (b) for students from the total sample who were not in the final sample, at each of the sixteen institutions in the principal conditions. These scores were compared by two-way analysis of variance, blocking on institution. The mean pretest score for those in the final sample (44.76) proved to be quite similar to that of the other students (46.56), and the difference between these two means was not significant ($F(1,756) = 1.38$). Therefore, the final sample was considered to be representative of the total group, and all analyses were based on the final sample.

The analysis did, however, reveal a significant effect of institutions ($F(15,756) = 4.89$, $p < .001$), indicating that there was a substantial amount of variation in average pretest score between institutions. Institutional means of the pretest scores, in percent correct, ranged from 36.62 to 59.12. Analyses to be presented below used pretest scores to control for group differences in initial English proficiency.

Analyses of Total Test Scores

Mean posttest scores for each of the sixteen principal institutions are presented in Table 3 on page 15. Note that the mean score for the disclosed posttest is presented first for each institution, regardless of the order in which the tests were presented. Standard errors are presented in parentheses below the corresponding means. (The standard error is equal to the standard deviation divided by the square root of N--number of students--on which the mean is based.) Standard errors rather than standard deviations are presented here since this statistic allows a visual impression of the statistical reliability of the means and the differences between means. Standard deviations for individual institutions ranged from 7.9 to 16.8, with average standard deviations over all sixteen institutions of 11.63 for the undisclosed posttest and 13.05 for the disclosed posttest.

The most striking aspect of the data is that the difference in scores was almost exclusively in the direction of greater scores on the disclosed than the undisclosed posttest. This was true for all but one institution. Since the basic difference in construction of these tests lay in the fact that items in one test had been disclosed to the

Table 3

Mean Posttest Scores for Each Institution (Standard Errors are in Parentheses)

6-form disclosure condition								12-form disclosure condition							
Disclosed posttest admin. first (Form A)				Disclosed posttest admin. second (Form B)				Disclosed posttest admin. first (Form A)				Disclosed posttest admin. second (Form B)			
Inst.	N	Score discl. test	Score undis. test	Inst.	N	Score discl. test	Score undis. test	Inst.	N	Score discl. test	Score undis. test	Inst.	N	Score discl. test	Score undis. test
A	16	52.56	49.07	E	37	55.92	46.85	I	28	58.70	56.10	M	24	59.86	55.62
		(2.99)	(2.56)			(2.30)	(2.05)			(2.03)	(1.84)			(2.69)	(2.31)
B	70	54.59	51.12	F	19	43.12	38.66	J	76	57.82	55.30	N	53	50.24	47.15
		(1.66)	(1.37)			(3.36)	(3.31)			(1.50)	(1.19)			(2.06)	(1.82)
C	42	57.88	53.43	G	65	54.93	46.54	K	18	55.09	55.54	O	8	49.85	46.72
		(2.59)	(1.92)			(1.53)	(1.31)			(2.11)	(2.08)			(3.15)	(3.46)
D	12	64.28	58.26	H	27	65.28	54.47	L	31	51.87	51.23	P	3	62.33	54.22
		(2.87)	(2.29)			(2.71)	(2.50)			(2.55)	(2.51)			(8.13)	(7.97)

students while items in the other had not, performance apparently was improved by disclosure of the items. The analysis described in the next section was performed to determine the magnitude and significance of the effect produced by item disclosure.

Analysis of covariance. The difference between disclosed and undisclosed posttest scores was computed for each student in the four experimental subgroups. An analysis of covariance was performed on these "disclosed-undisclosed difference scores" with pretest score as the covariate and two treatment factors: (a) number of disclosed forms (six vs. twelve), and (b) test order (disclosed test first as Form A, vs. second as Form B).[3] A mixed-effects covariance model was employed; since number of forms and test order were randomly assigned to institutions, these two factors were included as fixed effects and institutions as a random effect. In accord with this model, the error term in the analysis was based on variation between institutions within treatment groups. (An analysis of covariance was performed with sex of student as an additional factor. The effect of sex proved nonsignificant ($F(1,12) = 1.14$), indicating the appropriateness of combining the sexes in the principal analysis.)

In essence, analysis of covariance tests for effects of the experimental factors, with adjustment for group differences in pretest scores. The adjustment employs a common regression model in which the disclosed-undisclosed difference score is regressed on the pretest score. The regression slope of .042 obtained here indicated a significant ($p = < .03$) relationship between pretest score and difference score. A separate covariance analysis indicated that the regression slopes did not differ across the four experimental subgroups.

As the model assumes that the disclosed-undisclosed difference score depends both on effects of the experimental variables and pretest score, group means adjusted for pretest score are presented here. Adjusted group means were calculated by adding an adjustment factor to the raw group difference score. This adjustment factor was equal to the regression slope from the analysis of covariance multiplied by the difference between: (a) the group pretest score and (b) the mean pretest score for all groups combined. An example helps to demonstrate the procedure. For the group given six forms with disclosed posttest first,

---

[3]Although difference scores are known to have relatively low reliability, Overall and Woodward (1975, 1976) have shown that this is not a problem when the difference scores are being used as a basis for examining treatment effects, as was the case here. In fact, these authors show that the power of tests of significance is actually highest when the reliability of the difference scores is lowest--i.e., when the correlation between tests is high and, thus, the standard error of the difference and the reliability of the difference are low.

the mean disclosed-undisclosed difference score was 4.36. The mean pretest score for that group was 46.85, and the mean pretest score for all groups combined was 44.76. Hence,

Adjusted mean difference score = 4.36 + ,042 x (44.76 - 46.85) = 4.27.

Adjusted mean difference scores for the four experimental subgroups are presented in Table 4 on page 18.

The first result of interest was that the overall mean difference score of 4.60 was significant ($t(12) = 13.84$, $p < .001$).[4] Thus, there was a significant disclosure effect, as item disclosure increased scores on the TOEFL by 4.6 percentage points, averaged across experimental conditions. Significant main effects were also obtained for the number of disclosed forms ($F(1,12) = 22.84$, $p < .001$), and test order ($F(1,12) = 24.11$, $p < .001$), with no significant interaction between these factors. The effect of number of disclosed forms indicates that students receiving six disclosed forms had higher disclosed-undisclosed difference scores than did students receiving twelve disclosed forms. This is shown in the means at the bottom of Table 4. Before conclusions can be drawn about this effect it is necessary to examine the outcome of an additional analysis, presented below, which eliminates the influence of a confounding factor (see "Analysis for Items Not Covered in Class"). The test-order effect indicates that the mean difference score was greater for students administered the disclosed test second, as Form B, than for those administered this test first, as Form A. Analysis of data for the control institutions, presented below, helps to separate this test-order effect from the effect of disclosure.

Analysis of differences between institutions. To look for differences between institutions within each experimental subgroup, contrasts on the adjusted difference scores were calculated. Each institution within a subgroup was compared with each of the other three institutions in that subgroup, using a significance level appropriate for multiple comparisons. None of these contrasts reached significance at $p = .05$. (Only two of 24 possible tests even exceeded a t value of 2.00: $t(54) = 2.04$ for Institutions E vs. F; $t(44) = 2.57$ for Institutions F vs. H). In general, then, the variation among institutions was not greater than would be expected by chance.

---

[4]One-tailed comparisons were used to assess the significance of the overall disclosure effect and the significance of the disclosure effect for each experimental condition, since these effects were expected to be positive in all cases. T tests were used for these comparisons, which determined whether the mean disclosed-undisclosed difference score in each case was greater than zero. Other effects were assessed with two-tailed tests; they are presented as F tests as they are based on analyses of covariance.

Table 4

Adjusted Mean Disclosed-Undisclosed Difference Score
For Each Condition (Standard Errors are in Parentheses)

	6-form disclosure condition	12-form disclosure condition
Disclosed posttest administered first	4.27 (.65)	1.19 (.62)
Disclosed posttest administered second	8.33 (.64)	4.59 (.82)
Mean	6.30 (.46)	2.89 (.51)

Note: The mean difference scores without statistical adjustment are for the 6-form condition: 4.36 and 8.18, respectively, for the two test orders; for the 12-form condition they are 1.00 and 4.65, respectively.

Nevertheless, a dimension of difference among institutions that one might expect to relate to the disclosure effect is the amount of time the materials were available for study. Hence, a correlation was computed, for each of the four experimental subgroups, between (a) mean adjusted difference score and (b) days from distribution of disclosed forms to first posttest (see Table 2 on page 12). Institution was the unit of analysis. The average of these correlations across the four experimental subgroups was .00, indicating no relationship between these two variables.

Analysis of data from control institutions. Table 5 below presents the data for the four control institutions. These data suggest that Form A and Form B were relatively comparable in difficulty level. Nevertheless, on average there was a tendency toward higher scores for Form B than Form A and a tendency toward higher scores on the test administered second (Form B for Institutions Q and R, Form A for Institutions S and T) than the test given first. Therefore an additional analysis was performed using data from the control institutions to separate the effects of these factors from the disclosure effect. This analysis is presented in Appendix A. The results indicate that, when effects of the aforementioned factors were eliminated, the disclosure effect was significant for each of the four experimental subgroups and was relatively uniform across the two subgroups of the 6-form disclosure condition and across the two subgroups of the 12-form disclosure condition.

Table 5

Mean Posttest Scores for Each Control Institution
(Standard Errors are in Parentheses)

Posttests admin. in order Form A-Form B				Posttests admin. in order Form B-Form A			
Inst.	N	Score Form A	Score Form B	Inst.	N	Score Form A	Score Form B
Q	31	56.89	57.94	S	68	61.46	61.88
		(2.67)	(2.12)			(1.36)	(1.26)
R	33	45.89	48.02	T	7	57.13	56.50
		(1.85)	(1.90)			(3.51)	(4.51)

Analysis for Items Covered in Class

Class discussion was devoted to the items in Section I of one disclosed form and the items in Sections II and III of another form. Hence, the equivalent of one full form was covered in class, or a total of 150 items. The specific forms selected for class coverage were the same for all four institutions within a given experimental subgroup; they differed across subgroups, however, since each subgroup received a different set of forms. Since each disclosed form contributed about the same number of items to the disclosed posttest, about 1/6 of the items on the disclosed posttest were covered in class in the 6-form condition, whereas only about 1/12 of the items on the test were covered in the 12-form condition. For the latter condition, the number of items involved was too small, and the range of difficulty of these items too restricted, to yield an effective test of disclosure effects. Hence, analyses of disclosure effects for items covered in class were based on students in the 6-form disclosure condition only.

Computation of score. The method of computing the score for items covered in class is explained in detail in Appendix B. Briefly, for each subsection of each posttest the items covered in class, or "covered items", were identified. The percent correct for these items was computed for each subsection. Then, the score for a full section was computed as a weighted mean of subsection scores, where the weights were the total numbers of items contained in these subsections. The section scores were then averaged to yield a total percent correct.

The covered-item score for the disclosed posttest was based on items designated for coverage in class. To provide a basis for comparison, a score was also computed for the undisclosed posttest based on items designated for class coverage by students given the opposite test order (i.e., those given the other posttest form as the disclosed test). Thus, two scores were computed for each student, one based on items in the disclosed posttest and the other based on items in the undisclosed posttest. The difference between these two scores, averaged across test orders, provides an estimate of the effect produced by class coverage.

Analysis. A disclosed-undisclosed difference score was calculated for each student, based on the scores for the items covered in class. A one-way analysis of covariance was performed on these scores with test order as the independent variable and pretest score as the covariate. The overall mean difference score was 11.80, indicating that class coverage significantly increased performance on the covered items by 11.8 percentage points ($t(6) = 20.07$, $p < .001$). The test-order effect was not significant.

Analysis for Items Not Covered in Class

Recall that each student was asked to bring two of the disclosed forms to class. Items not contained in either form brought to class are termed "noncovered items." Posttest scores for noncovered items were computed using a method analogous to that described in the previous section. That is, the percent correct for noncovered items was determined for each subsection, and these figures were combined into section scores with each subsection weighted by the total number of items in it. The noncovered item score was a simple average of the three section scores.[5]

For each student the difference between disclosed and undisclosed posttests was calculated for noncovered items. The data were analyzed by two-way analysis of covariance with number of disclosed forms and test order as factors and pretest score as the covariate. The first effect of interest was that the overall mean difference score, 3.19, was significantly different from zero ($t(12) = 7.22$, $p < .001$). Thus, item disclosure increased the scores for noncovered items an average of about 3.2 percentage points. A significant effect of number of forms ($F(1,12) = 6.19$, $p < .05$) indicates that students given six disclosed forms obtained a higher mean difference score (4.39) than did those given twelve forms (2.00). Nevertheless, the difference score was significantly different from zero for both the 6-form condition ($t(12) = 7.43$, $p < .001$) and the 12-form condition ($t(12) = 3.09$, $p < .01$), showing that even for students given twelve disclosed forms the effect of item disclosure was statistically reliable.

It was noted above that the difference in total test scores between the 6- and 12-form conditions could have been due to confounding. Specifically, for the 6-form condition, 1/6 of the items on the disclosed posttest were covered in class, whereas for the 12-form condition only 1/12 of the items on the disclosed posttest were covered in class. Hence, a difference in total scores for these conditions could have been due simply to the fact that students in the 6-form condition received in-class exposure to more of the items on the test than did students in the 12-form condition. The noncovered-item analysis presented here, however, is not subject to such confounding and demonstrates that the disclosure effect was indeed greater for students given six disclosed forms than for those given twelve forms.

---

[5]All items in the two forms brought to class were excluded in deriving this score rather than just the items actually discussed in class. It was reasoned that the instructor's request to have both test forms on hand could have induced the students to study all items in both forms. The question under study is whether the students examined disclosed items on their own initiative, and this question is best addressed by investigating effects only for items in test forms the students were not requested to keep on hand.

The principal analyses are those discussed in the preceding section, as these analyses addressed the major questions at issue in this study. In an effort to obtain further information about the disclosure effect and the conditions under which it is observed, supplementary analyses were also performed, as described in the sections to follow.

Analysis by Language Group

For this analysis the data were grouped according to the students' native language, as listed above in the Method section. With the exception of French, these languages fell into three logical groups, which accounted for 85% of the final sample: (a) Spanish and Portuguese (39% of sample), (b) Arabic and Farsi (26%), and (c) Japanese, Chinese, Vietnamese, Thai, and Korean (20%). For the first language group the number of students per institution ranged from 2 to 27; for the second group, the number ranged from 0 (at one institution) to 40; and for the third group, from 0 (at two institutions) to 20. Disclosed-undisclosed difference scores for each group were analyzed by a two-way analysis of covariance with number of forms and test order as factors and pretest score as the covariate.

The analyses revealed significant disclosure effects of 3.37 percentage points for the Spanish-Portuguese group ($p < .01$); 4.31 percentage points for for the Arabic-Farsi group ($p < .05$); and 6.18 percentage points for the Asian group ($p < .01$). Therefore, scores for students in every language group were affected by item disclosure.

As in the overall analysis presented earlier, the effect of test order was significant for each group. The effect of number of disclosed forms, while in the same direction for each group as that observed in the overall analysis, was not significant. (This was due to the fact that the error terms in the analyses for individual language groups were considerably higher than those in the overall analysis, resulting from the relatively small numbers of observations involved.) An overall comparison indicated that the disclosure effect was not significantly related to the students' native language ($F < 1$).

Analysis by Test Section

To better understand the disclosure effect, data for each subsection of the test were analyzed separately. For this analysis it was not appropriate to use a score based on all items since the numbers of items covered in class were disproportionately large for some subsections of the test. Hence, this analysis was based on items not covered in class, or "noncovered items." An additional analysis for items covered in class is also reported.

The percent correct for noncovered items was computed for each of the seven subsections of the test, and the difference between disclosed and undisclosed posttest scores for noncovered items was calculated for each subsection. Difference scores for each subsection were analyzed by two-way analysis of covariance with number of forms and test order as factors and the corresponding pretest subsection score as the covariate.

The disclosure effects by subsections were: Section I: 3.89, 4.59, and 3.40 percentage points; Section II: 3.46 and 2.21; and Section III: 4.67 and 1.51. (Initial performance levels in the undisclosed test averaged across conditions were, for the seven subsections, respectively: 59.1, 64.8, 58.1, 46.0, 43.7, 45.7, and 49.6.) The disclosure effect was significant in all cases ($t(12) \geq 1.82$, $p < .05$), showing that item disclosure affected the scores for every item type. (The effect of number of disclosed forms was not significant for any subsection; the effect of test order was significant in four cases--Subsections I-1, II-2, III-1, and III-2.)

The mean disclosure effects listed above appeared to vary somewhat across subsections of the test. The largest effects were observed for Vocabulary (Section III-1) and Listening Comprehension, Dialogues (I-2), while the smallest effects were observed for Reading Comprehension (III) and Written Expression (II-2). The effects for the other three subsections fell between these extremes. While these differences are suggestive, the variation among these seven scores was not significantly greater than that expected by chance ($F(6,84) = 1.07$). Furthermore, a logical ordering of subsections is not readily discernible in these means. Although the data may suggest hypotheses to be tested in further research, for the present it cannot be concluded that disclosure produces different effects for different item types. Of primary importance is the fact that no item type was insensitive to the effects of disclosure.

Difference scores for items that were covered in class, or "covered items," were also analyzed. As before, this analysis was performed for the 6-form condition only. The data were combined into section scores for each of Sections I and II since there were very few covered items in each subsection. For Section III, however, the data were analyzed for each subsection separately due to the structural dissimilarity of the item types in these two subsections (Vocabulary and Reading Comprehension). Disclosed-undisclosed difference scores were submitted to analysis of covariance with test order as a factor and the score on the corresponding section or subsection of the pretest as the covariate.

The disclosure effects for the four parts of the test were: Section I: 9.2 percentage points; Section II: 11.9; Subsection III-1: 14.8; Subsection III-2: 11.5. Each of these effects was significant ($t(6) \geq 4.56$, $p < .01$), indicating that item disclosure influenced performance for every one of these item types. Variation in the disclosure effect across Sections I, II, III-1 and III-2, however, was not significant ($F(2,18 = 1.36$). The effect of test order was significant for Section II and Subsection III-2 ($F(1, 6) \geq 13.39$, $p < .01$).

Analysis of Student Questionnaire Data

All participating students were asked to respond to a brief questionnaire at the end of the study. This questionnaire was designed to elicit information about the amount of effort devoted to studying the disclosed materials. It should be kept in mind that these data are based on a self-report procedure that tests students' memory of what they had done and, as such, are potentially biased. Nevertheless, these data provide a rough guide to the students' study activities. Returns were obtained from fourteen institutions--all eight institutions in the 6-form condition, and six institutions in the 12-form condition (three for each test order). The analyses to be reported here are based on the responses of all students present in class when the questionnaires were filled out. Instructors were asked to read the questions aloud as the students read them on their questionnaires.

Two questions asked about the time spent studying the disclosed materials:

Question 1. How much time did you spend listening to the tapes outside of class?

Question 2. How much time did you spend reading the booklets outside of class?

Many students gave responses that were uncodable, such as "many times." These students were excluded from the principal analyses. The analyses for Questions 1 and 2 were based on 506 of the 645 students responding to the questionnaire. Institution averages ranged from 1.0 hours to 5.8 hours for listening time, and .8 hours to 9.4 hours for reading time. The mean listening time across institutions was 3.37 hours for the 6-form condition and 4.01 hours for the 12-form condition; mean reading times for these two conditions were 5.43 hours and 6.01 hours, respectively. T tests comparing the 6- and 12-form conditions (with institution as the unit of analysis) showed the difference to be nonsignificant for both listening time and reading time.[6]

---

[6]Some students responded that they spent a certain number of hours per day or per week. Although these responses could not be considered in the main analysis, since they require subjective interpretation, a supplementary analysis was performed including these 38 students. The amount of time indicated by such a student was multiplied by the number of days or weeks between distribution of the disclosed forms and the posttests at that student's institution (see Table 2 on page 12). Combining data for these students with data for the 506 students in the above analysis, mean listening times across institutions were 5.43 hours for the 6-form condition and 5.53 hours for the 12-form condition; mean reading times were 7.16 hours and 8.05 hours, respectively, for these two conditions. The average times were not significantly different for the two conditions.

Two additional questions asked about the amount of material covered:

Question 3a. How many tapes did you listen to? _____

Question 3b. How much of each tape did you listen to, on the average?

1/4 1/2 3/4 All

(The instructor explained that Question 3b pertained to the tapes indicated in response to Question 3a.)

Question 4a. How many booklets did you read? _____

Question 4b. How much of each booklet did you read, on the average?

1/4 1/2 3/4 All

(The instructor repeated an instruction similar to that for Question 3b.)

Data were excluded for those students who failed to provide a response to both parts of both questions. Analyses for Questions 3 and 4 were based on 567 out of 645 total questionnaires. For each of Questions 3 and 4 the responses to parts a and b were multiplied together to yield the amount of material studied, expressed in number of full tapes or full booklets covered.

The average numbers of tapes covered were 2.78 for the 6-form condition and 3.62 for the 12-form condition. The average numbers of booklets covered were 2.61 and 4.30 for these two conditions. The difference between conditions was significant for number of booklets (t (12) = 2.73, $p < .05$) but not for number of tapes.

It might have been desirable to ascertain whether the disclosure effect was related to the amount of effort that students said they devoted to working with the materials. Unfortunately, limitations inherent in the data prevented assessment of such a relationship (or assessment of a relationship with instructor questionnaire data to be presented below). Nevertheless, the student questionnaire data are important in demonstrating that the amount of time spent by the students did not increase with an increase in the amount of material available for study.

Instructor Questionnaire Data--Overview

Each instructor involved in class coverage of the disclosed materials was asked to respond to a questionnaire at the end of the study, to obtain information about the amount of material covered and about the nature of the class-coverage experience. The analysis of the questionnaire data is presented in Appendix C. Among the general impressions that emerge from this analysis is that instructors conformed quite closely to

the specifications set forth at the outset of the study. They devoted an average of about five hours to class coverage. They had the students read (listen to) almost all of the items in the designated forms and practically no other items. On occasion, an instructor discussed non-designated items in class or discussed the disclosed materials with students outside of class, but this occurred infrequently.

Regarding qualitative aspects, the instructors generally reported that they had the class read or listen to each item individually rather than having them read sections or the entire test at a time. Discussion was generated by both students and instructors, with a relatively even balance between items selected for discussion by instructors and those selected by students. When asked about the students' interest in discussing the various sections of the test, the instructors noted that the students were primarily interested in Structure and Written Expression and Vocabulary, and they were less interested in Listening Comprehension and Reading Comprehension. The students' interests thus seemed to relate to the distinction between discrete and integrative items types. Discrete items are those requiring focus on specified linguistic features of English, whereas integrative items are presented in a larger context and depend on an interplay of linguistic, situational and discourse features. Although no further data exist to support firm conclusions regarding this distinction, students in this study did appear to be most interested in discussing those items with a central focus which they might expect to remember more easily.

DISCUSSION

Overall Disclosure Effect

Performance on the TOEFL was clearly affected by item disclosure. When items were made available to the students for a few weeks prior to administration of the test, the students studied at least some of these items and increased their scores as a result. This is shown in significantly greater scores obtained on the TOEFL containing items that had been made available beforehand than on the test containing all new items.

There are two categories of effect that can result from test preparation activities with disclosed items. The first is a general learning effect, i.e., an improvement in performance resulting from experience with the TOEFL in general. The second is a specific recall effect, i.e., an increase in performance due to recall of the specific questions and answers encountered in the test. It is this second effect with which the present study is concerned. Any general learning about the TOEFL that may have occurred would have affected performance in the two posttests equally. However, a tendency to recall specific items from the disclosed forms would have increased scores only for the posttest containing those items. The observed difference in posttest scores, therefore, shows that the students studied the disclosed forms and recalled specific questions and answers from those forms upon encountering them on the test.

It is believed that a disclosure effect, produced experimentally here, would also be observed in an operational test situation if the test included disclosed items. This belief rests on the assumption that the present situation simulates conditions that would occur in reality, at least for many TOEFL candidates. There are good reasons to accept this assumption. Special schools provide instruction in taking the TOEFL, and many English language programs have instituted sessions or classes in test preparation. Also, students often establish their own study groups for this purpose. It is reasonable to expect that disclosed TOEFL forms will become included among the study materials for such classes or groups, not necessarily because instructors or students intend to boost scores through item memorization but simply because publicly available TOEFL forms are seen as useful practice tests. Assuming that such forms become part of the study materials, they will undoubtedly be made highly accessible to the students, with copies distributed widely and selected forms used for tutorial purposes. The students' motivation to study such forms should be relatively high, since the incentive to perform well should be at least as great for an operational TOEFL as for the experimental tests used in this study.

The average disclosure effect--the difference between scores for the disclosed and undisclosed posttests--was 4.6 percentage points

(6.3 percentage points for the 6-form condition and 2.9 percentage points for the 12-form condition; more about the difference between conditions below). This effect was equal to about one-third of a standard deviation in test scores and can be expressed as a difference of about 18 points on the TOEFL scale (see Appendix D).

The import of the disclosure effect can best be understood, however, with reference to the estimated number of items for which the students knew the correct answers. While the average score on the undisclosed posttest was around 50%, the percentage of items for which the students knew the correct answers can be assumed to have been smaller than that, since many questions presumably were answered correctly by guessing. The lesser percentage provides the more appropriate baseline against which to compare the increase due to disclosure, since the increase represents the additional items for which the students knew the correct answers. The point is best demonstrated by a concrete example. Consider a test with 72 items and 4 alternatives per item. (This test size is chosen for convenience; the numerical relationships hold for tests of any size). Suppose that, on the average, examinees know the answers to 24 items and are totally guessing on the other 48. They are expected to answer 12 items correctly by chance for a total of 36 correct responses, or 50% correct. (A score of 50% correct thus reflects knowledge of 33% of the answers.) If examinees were to increase the number of answers known by 4, they would be expected to obtain a score of 39 correct (28 known answers plus 11 correct by chance) or 54.2% correct. Hence, an increase in score of just under 4 1/2 percentage points in this case represents an increase in number of known answers by 1/6.[7] This example provides a rough indication of the increase in number of answers known as a result of item disclosure. Of course, the precise increase cannot be determined, as it is impossible to know the extent to which students had partial knowledge of items (i.e., could eliminate one or two alternatives). It is also impossible to determine whether the effect of studying disclosed test forms was to increase the number of fully known items or to increase the number of partially known items, or both. The general point, however, is that an increase in score of a few percentage points represents a relatively sizeable increase in number of items for which students knew some information.

---

[7]Perhaps not all items are either fully known or subject to total guessing. Consider a situation in which, on the average, the answers to 10 out of 46 items are known, and the remaining 36 items are equally distributed among those for which students can eliminate (a) two alternatives, (b) one alternative, and (c) no alternatives. Simple arithmetic shows that the students' expected average score in this case is 50%. If, through studying disclosed items, the students were to increase by 2 the number of items falling into each of the "partially known" categories (a and b) and the "fully known" category, then their expected average score would be 54.7% correct. Thus, an increase in score of just over 4 1/2 percentage points in this case represents about a 1/6 increase in number of known or partially known answers.

Effects of Class Coverage and Independent Study

The effect of disclosure discussed above represents the combined result for items that were covered in class and items that were available for independent study. This result thus indicates the overall effect that might be expected if a portion of the disclosed items were subject to direct instruction and the remainder were available for students to study on their own initiative. To understand the basis for this effect, it is useful to examine separately the impact of class coverage and the effect of independent study. Toward this end, two followup analyses were performed, one for items that were covered in class and the other for items that were not covered.

The analysis for items covered in class was performed for the 6-form condition only, due to the restricted item sample in the 12-form condition. The results showed a disclosure effect of 11.8 percentage points for these items, a quite marked increase. It is perhaps not surprising that a relatively large effect was observed. The class coverage ensured exposure to these items, and most of them were subject to careful scrutiny via the class discussion. It is not claimed that an increase of this size would necessarily be observed for total test score if all disclosed items were to be covered in class. Possibly such an effect would be observed if instructors were able to devote sufficient time to coverage of all items, but this is a question that must remain for further research. The results obtained here show the kind of effect that can be expected for a subset of items that are subject to careful examination.

The effect for items covered in class, while quite pronounced, was by no means solely responsible for the overall disclosure effect. A significant effect was shown also for the items that were not in the forms covered in class. The average effect for these items was 3.2 percentage points--4.4 in the 6-form condition, and 2.0 in the 12-form condition. The students apparently engaged in independent study of the disclosed forms that were not covered in class. The questionnaire data also support this conclusion. The students reported having averaged approximately three to four hours listening to the tapes and five to six hours reading the disclosed forms (slightly higher averages were obtained when ambiguous responses were counted). The questionnaire data admittedly are subject to self-report bias and memory limitations. Nevertheless, these data do show that the students devoted some effort to independent study of the materials not covered in class.

The students appeared, therefore, to be generally motivated to study the disclosed forms on their own initiative. Of course, the class experience may have enhanced the students' interest in test preparation and may have implied faculty encouragement to study the disclosed materials. Nevertheless, the students were under no compulsion to study these materials. They had been told at the outset that these forms were being made available for their use and that the forms contained items that would appear on the special TOEFL, but they were not required to study

them. It seems reasonable to conclude, therefore, that students are motivated to study disclosed test forms without any specific assignment to do so. Whether their motivation derives from a desire to gain general experience with the TOEFL or from an intention to memorize specific items they expect to encounter, the practical implications are the same. When disclosed TOEFL forms are made available, students can be expected to study those forms and remember specific items when encountering them on a subsequent TOEFL.

Variation in Effect Due to Size of Disclosed Item Pool

The disclosure effect of 4.6 percentage points is an average. In fact the magnitude of the disclosure effect appears to vary with the number of items in the pool from which those in the test were drawn. A disclosure effect of 6.3 percentage points was observed for students given six disclosed forms (or 900 items), whereas a significantly lower effect of 2.9 percentage points was observed for those given twelve forms (1800 items). For items not covered in class, on which a direct comparison is more appropriately based, the effects for these two conditions--4.4 and 2.0 percentage points, respectively--were also significantly different. The data thus suggest a general principal, applicable at least within limits of the present experimental conditions: If students must cover a relatively large number of test forms in order to be exposed to all items that will appear on a later test, they are less likely to benefit from disclosure than if they need cover a smaller number of forms.

It is reasonable to speculate that further enlargement of the disclosed item pool would bring a further reduction in the effect of disclosure. A reason for such an hypothesis is that, with the present increase in item-pool size, the disclosure effect was diminished by more than half, yet it remained statistically significant (for both total score and items not covered in class) leaving considerable room for further reduction. Also, one might logically expect the disclosure effect to approach zero as the item-pool size approaches infinity, since increasingly larger item pools would be associated with increasingly smaller amounts of study time per item. The results observed in this study, then, are likely part of a general function relating increases in item-pool size to decreases in the disclosure effect.

The item pool in the 12-form condition, while producing a significant disclosure effect, was large enough that many students who received twelve forms seemed to be overwhelmed, according to comments offered by many instructors. Further, the student questionnaire data suggest limitations in students' ability to devote maximally effective study time to an item pool of this size. The reported amount of time spent covering the disclosed forms was only slightly but not significantly greater for the 12-form than the 6-form condition; yet the students read through significantly more material in the former condition (the number of tapes heard, however, did not differ significantly across conditions).

The data thus fit the pattern that would be expected if there were a limit in amount of time students will devote to study of disclosed materials, regardless of the number of forms to be covered or the number of forms they attempt to cover. If there is indeed such a limit, it is understandable that students would feel increasingly overwhelmed and that the disclosure effect would gradually decrease as the item-pool size were increased; students would attempt to spread about the same amount of study time over more and more material, allowing less careful study per item.

The amount of study time may, of course, depend on the period of time the materials are available for study. In the present experiment the disclosed forms were available for four or five weeks in most cases and six weeks at the most. If disclosed items were to be made available for a much longer time, students would be able to study a large item pool more carefully. Whether they would actually do so is an open question as there may be a limit to how early students would begin preparation for a TOEFL in earnest--perhaps only a few weeks before the test as in the present situation. Nevertheless, the longer disclosed items are available, the greater is students' opportunity to study them--and thus, the greater is the expected effect of disclosure on TOEFL performance.

Other Results

It is important, finally, to note that the disclosure effect was relatively uninfluenced by several factors. The results did not differ significantly across institutions within subgroups, despite differences in the nature of the institutions and the procedures by which they carried out the experiment. Also, the disclosure effect was significant for each of the three principal language groups in the sample and did not differ among them. Further, the effect was significant for each of the seven subsections of the test, with no clear differences among them. In thus showing that the effect is not unique to any single group of students or type of item, these results demonstrate that the effect of item disclosure on TOEFL performance is a robust phenomenon.

Implications for the TOEFL Program

As explained in the introduction, a new law in New York State and pending legislation in other states had raised potential problems, particularly for the TOEFL Institutional testing program. Prior to initiation of this study, TOEFL forms used at International or Special Center administrations

were provided to institutions for internal use. Under the requirements of the new legislation, however, many forms previously eligible for use in the Institutional program could no longer be used since they would have been disclosed. An alternative under consideration was not to reuse complete forms but to combine items from previously used forms into new ones for administration in the Institutional program. The appropriateness of such a plan would depend on the effect of item disclosure on test performance; hence, the present study was undertaken to examine this effect.

Since this study began, amendments to the New York State law and the delay in further legislative action in other states have reduced the immediate problem for the TOEFL Institutional program. It has been possible to select forms for Institutional use that do not contain any items that have appeared in disclosed tests. Should more restrictive requirements come into effect, however, the evidence collected in this study will play an important role in determining whether procedures involving reuse of items can be considered.

The study has shown that access to disclosed test items does, indeed, increase scores on a TOEFL containing those items. Even when students had as many as twelve forms to cover, or 1800 items, there was a small but significant increase in performance. It is difficult to estimate the effect that would be produced if items were disclosed over a long period of time in an operational test situation. However, it is reasonable to assume that the effect would increase with an increase in the period of time that items are available. Furthermore, students' motivation to practice with available items should be high; the TOEFL, as an English proficiency test, is particularly susceptible to practice, and there are pressures for foreign students to present acceptable TOEFL scores in order to begin academic work at many colleges and universities. Considering these factors, the appropriateness of reusing disclosed test items appears questionable. Circumstances may change as more and more TOEFL forms become disclosed, given that an increase in the pool of disclosed items apparently brings a decrease in magnitude of the disclosure effect. However, for the near future at least--when a limited number of forms will have been made available--it appears advisable not to reuse disclosed TOEFL items for institutional use but to employ alternative test development procedures.

REFERENCES

Overall, J. E., & Woodward, J. A. Unreliability of difference scores: A paradox for measurement of change. Psychological Bulletin, 1975, 82, 85-86.

Overall, J. E., & Woodward, J. A. Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. Psychological Bulletin, 1976, 83, 776-777.

APPENDIX A

Analysis of Data from Control Institutions

Data for the control institutions (Table 5 on page 19) showed that, on average, the score for Form B was higher than that for Form A, and the score for the test administered second (Form B for Institutions Q and R, Form A for Institutions S and T) was higher than the score for the test given first. Therefore, for students in the experimental groups there were actually three factors operating to determine the difference between disclosed and undisclosed posttest scores. For students given the disclosed posttest second, Form B, the difference in favor of the disclosed posttest was due to the effect of item disclosure, augmented by: (a) the positive effect of taking the disclosed posttest second, and (b) the positive effect of taking Form B as the disclosed posttest. For students given the disclosed test first, Form A, the disclosed-undisclosed difference was due to the effect of item disclosure diminished by: (a) the negative effect of taking the disclosed posttest first, and (b) the negative effect of taking Form A as the disclosed posttest.

An analysis was performed to estimate the effect of item exposure independent of the effects of the other two factors mentioned above (i.e., first test vs. second test, Form A vs. Form B). For this analysis, the two control institutions (Q and R) that received the tests in order Form A-Form B yield the appropriate data, since the difference between posttests for these institutions represents the combined effects of these two factors. The average score on Form B, presented second, was 1.59 percentage points higher than the average score on Form A, presented first. This difference was adjusted to account for the fact that the mean pretest score for these two control institutions taken together was 46.02, in contrast with 44.76 for the institutions in the principal conditions. The formula for this adjustment was:

Adjusted mean difference = 1.59 + .042 x (44.76 - 46.02) = 1.54

For the "disclosed posttest first" condition, 1.54 was added to the means in Table 4 on page 18, whereas, for the "disclosed posttest second" condition, 1.54 was subtracted. The resulting figures indicate the expected gain due to item disclosure for each condition controlling for effects due to test form and order of presentation. The results were as follows:

6-form disclosure condition

disclosed posttest first	4.27 + 1.54 = 5.81
disclosed posttest second	8.33 - 1.54 = 6.79

12-form disclosure condition

disclosed posttest first	1.19 + 1.54 = 2.73
disclosed posttest second	4.59 - 1.54 = 3.05

Thus, the expected gain due to disclosure was between 5.8 and 6.8 percentage points for the 6-form condition and between 2.7 and 3.1 percentage points for the 12-form condition. All of these expected gains were significant ($t(12) \geq 3.00$, $p < .01$).

It will be noted that the means at the bottom of Table 5 (6.30 and 2.89) also provide estimates of the disclosure effect for the 6-form and 12-form conditions with the effect of test form and order averaged out. The advantage of the analysis just discussed is that it demonstrates, using independent data from a control sample, that the magnitude of the expected disclosure effect was relatively consistent across both groups in the 6-form disclosure condition and across both groups in the 12-form condition.

APPENDIX B

Computation of Score for Items Covered in Class

As was true of scores for the whole test, a student's score for items covered in class was calculated by taking the score for each of the three sections and computing a simple average of these scores. The method of computing the score for each section, however, involves special considerations and requires detailed explanation.

For each subsection of each posttest, the items that had been covered in class, or "covered items," were identified. The percent correct for covered items was computed for each subsection. The score for a main section was then computed as weighted average of subsection scores, the weights consisting of the total numbers of items contained in those subsections. An example illustrates the procedure. Section I of the TOEFL is divided into three subsections containing 20, 15, and 15 items, respectively. In each of the posttests used in this study the numbers of covered items in these three subsections were 3, 2, and 5, respectively. Assume that, of these items, a student correctly answered 2, 1, and 2 items in these three respective subsections. The student's subsection scores for covered items, expressed as percentages, would be 67%, 50%, and 40%. The Section I score for covered items would then be the weighted average of these three scores:

$$\frac{(20 \times 67\%) + (15 \times 50\%) + (15 \times 40\%)}{50} = 54\%$$

The reasons for using this method of scoring are best conveyed by contrasting the method with possible alternatives. One alternative would be to identify all covered items in an entire section and sum the correct responses to these items. The problem with this approach lies in the disparity in the proportion of items from each subsection that were covered in class. For example, in Section I of each posttest, 3 of the 20 items in the first subsection were covered in class, while 5 of the 15 items in the third subsection were covered. This disparity resulted from the fact that questions in the third subsection occur in clusters. If the section score were the simple sum of all correct responses to covered items, performance in the third subsection would contribute disproportionately to the section score. It is necessary, therefore, to derive scores for individual subsections and combine them in order to avoid disproportionate representation of subsections.

To combine subsection scores weighted averaging is more desirable than simple averaging, as the former method maintains the original relationship among subsections. For example, consider Section II, whose two subsections contain 15 and 25 items. In a standard TOEFL, performance in the first subsection contributes 15/40 or 38% of the Section II score, and performance in the second subsection contributes 25/40 or 62%. When scores for these two item types are computed, whether based on all items in these subsections or a portion of those items, it is best to combine the two scores in a way that preserves this relationship.

APPENDIX C

Analysis of Instructor Questionnaire Data

Each instructor involved in class coverage of the disclosed materials was asked to respond to a questionnaire at the end of the study. Questionnaires were returned by fourteen of the sixteen institutions. For eight of these fourteen institutions, all instructors responded. The numbers of classes represented in these cases were 2, 2, 2, 2, 3, 4, 5, and 9. For the other six institutions the numbers of classes for which complete data were available, as a proportion of the number of classes participating, were: 7/8, 9/12, 3/5, 5/9, 2/4, and 2/4. The numbers of classes are given here rather than the numbers of instructors since team teaching was employed in many instances, and the class was the most logical unit of analysis. At five institutions responding to the questionnaire different instructors covered different sections of the test, while at the other nine institutions a single class instructor covered all parts of the test. In the analyses to be presented here, the data are examined only for the 57 classes for which complete data are available.

The first few questions pertained to the amount of time devoted and amount of material covered. For each of these questions the average response was computed across classes at each institution. (A simple average was derived, rather than an average weighted by number of students per class, since the class sizes were relatively uniform; the ratio of smallest to largest class size, averaged across institutions, was .78.)

Question 1: How many hours did you devote to class coverage of the disclosed materials?

The amount of time spent per class ranged from 3.3 hours to 10 hours. These figures were averaged across classes within each institution, and the average of the institution means was computed within experimental conditions. These means were 5.6 hours for the 6-form condition and 4.7 hours for the 12-form condition. Thus, the instructors devoted approximately 5 hours to class coverage, on the average. The difference between conditions was not significant, according to a *t* test with institution as the unit of analysis.

Question 2a: In the form designated for coverage of Section I, about how many of the 50 items in that section did your students listen to in class?

Questions 2b, 2c, and 2d: Same as Question 2a, for Sections II, III-1, and III-2, except that the phrase "listen to" was replaced by the word "read."

The average numbers of items covered in the 6-form condition were: 46, 39, 27, and 23 for Sections I, II, III-1, and III-2; for the 12-form condition these figures were 48, 39, 29, and 25. Thus, practically all items in the designated forms were read (heard) in class.

Question 3: About how many of the items mentioned in [Question 2] were discussed in class? Section I ___ Section II ___ Section III-1 ___ Section III-2 ___. (Note that the instructions to teachers at the outset of the study had drawn a distinction between having the students read (listen to) items in class and engaging in class discussion of them.)

The average numbers of items covered in the 6-form condition were 36, 35, 21, and 18 for the four parts of the test specified; for the 12-form condition these figures were 42, 36, 28, and 25. Thus, a large percentage of the items read (heard) in class were also discussed.

Question 4: In class, did the students read (listen to) items in forms other than those designated for class coverage? If so, about how many items?

Only one instructor responded "yes" to this question, noting that 10 additional items had been covered.

Question 5: Was there class discussion of items in forms other than those designated for class coverage? If so, about how many items?

Five instructors answered "yes" to this question, with four specifying the number of items discussed; the numbers ranged from 5 to 30 items, with an average of 15 items.

Question 6: Did you spend time discussing these materials with students outside of class? If so, about how much time?

Ten instructors answered "yes" to this question, with nine specifying the amount of time spent; the times ranged from 5 minutes to 1 1/2 hours, with an average of 30 minutes.

Three questions asked for qualitative information about the class coverage. For questions 7 and 8, the sample of instructors was the same as that described above, and the focus was on the class as the unit of analysis. Hence, where an instructor taught two classes (generally filling out one questionnaire for each class) his or her answers for each class were treated as two separate observations. Where a class was taught by a team, the modal response of the team members was taken as the response for that class.

Question 7: Did the students

a) read (and listen to) the whole test then discuss it

b) read (listen to) part of the test then discuss it (please specify roughly how many items were covered at a time)

c) read (listen to) and discuss each item in turn

d) other. Please specify.

The dominant response to this question was letter "c." The numbers of classes for which responses a, b, c, and d were given were 3, 8, 35, and 9, respectively. For 2 classes no response was given. Of those who responded "d" ("other"), a combination of these methods was most often indicated (although no clear pattern could be discerned).

Question 8: How were items selected for discussion in class?

a) by students' questions

b) through instructor initiation

c) both. Please specify proportion of items discussed in response to students' questions.

The dominant response here was "c." Instructors of 4 classes responded "a," 18 classes "b," and 35 classes "c." Where "c" was the response, the average proportion of items discussed in response to students' questions was .44 (averaged across the 28 classes for which a proportion was indicated).

For the next question, responses were examined only for 31 instructors who taught the whole test and who provided complete responses.

Question 9: How interested were the students in discussing each of the following types of items? (Check one for each type of item.)

Here the instructor was to check one of four alternatives for each of sections I, II, III-1, and III-2: "much interest," "some interest," "little interest," "no interest."

The numbers of persons checking these four alternatives, respectively, were: Section I: 11, 13, 3, 4; Section II: 22, 7, 2, 0; Section III-1: 14, 12, 5, 0; Section III-2: 3, 19, 5, 4. Thus, for Section II (Structure and Written Expression) and Subsection III-1 (Vocabulary) the modal response was "much interest," while for Section I (Listening Comprehension) and Subsection III-2 (Reading Comprehension) the modal response was "some interest," with some instructors responding "no interest."

APPENDIX D

Calculation of Scaled-Score Estimate of Disclosure Effect

Although scaled scores were not available for the posttests, it was possible to obtain an estimate of the disclosure effect in terms of the TOEFL scale. The method to be used for this purpose makes use of two facts: (a) the disclosure effect can be expressed as a proportion of the standard deviation (SD) of raw scores on the undisclosed posttest, and (b) data are available indicating the typical ratio of raw-score SD to scaled-score SD on the TOEFL.

First, regarding point "a" above, the average difference between disclosed and undisclosed posttests was computed for each of the three test sections. The difference for Section I was found to be .31 times the SD for Section I of the undisclosed posttest; for Sections II and III the proportions were .40 and .32.

Next, regarding point "b", data were examined for the International and Special Center administrations of the TOEFL between May and October, 1980 (three of each type). The data from domestic test centers only were considered, which yielded an average of 7154 examinees per administration. From these data the average ratios of raw-score SD to scaled score SD for the three test sections were calculated to be .78, 1.12, and .73.

Then, for the undisclosed posttest in this study, raw-score SDs were multiplied by these ratios to obtain estimates of what the scaled-score SDs would be if conversion to scaled scores were possible: 6.11, 5.54, and 5.61. These estimates were multiplied by the proportions indicated for point "a" above (.31, .40, and .32) to yield scaled-score estimates of the disclosure effect: 1.89, 2.21, and 2.47 units for the three test sections.

One additional correction was needed to account for the restricted range of proficiency levels in the present sample. The restricted range is seen in the fact that, for the undisclosed posttest, the SDs for the three test sections were, respectively, .95, .76, and .78 times as great as the average section SDs for the TOEFLs given in the abovementioned International and Special Center administrations. To make the necessary correction, these proportions were multiplied by the scaled-score estimates of the disclosure effect. The resulting disclosure effects per section were 1.79, 1.68, and 1.93 units. The mean of these figures multiplied by 10 equals 18.0, indicating that the overall disclosure effect (i.e., the average effect across experimental conditions) can be expressed as an estimated 18 units on the TOEFL scale.

The method used here does not correct for factors such as a possible difference between the mean proficiency level of the present sample and that of the typical International/Special Center sample. In the absence of actual scaling data, however, this analysis at least gives an index, possibly biased, of the magnitude of the disclosure effect in terms of the TOEFL scale.

TOEFL
Research Reports

The Performance of Native Speakers of English on the Test of English as a Foreign Language: Clark, John L.D. Report 1. November 1977.

Discusses the results of the administration of TOEFL to native speakers of English just prior to their graduation from a college-preparatory high school program. Total test score distributions were highly negatively skewed, reinforcing findings of earlier studies that TOEFL is not psychometrically appropriate for discriminating among native speakers of English with respect to English language competence.

An Evaluation of Alternative Item Formats for Testing English as a Foreign Language: Pike, Lewis W. Report 2. June 1979.

Describes an extensive research study conducted from 1972 to 1974 that was designed to explore possible changes in the format and content of TOEFL. Questions of validation, criterion selection, and content specifications were investigated. The report includes the results of these findings and discusses the implications for TOEFL content specifications and internal structure. This study contributed to the restructuring of TOEFL beginning in 1976.

The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests: Angelis, Paul J.; Swinton, Spencer S.; and Cowell, William R. Report 3. October 1979.

Gives the results of a study in which 400 graduate and undergraduate applicants took TOEFL, the GRE Verbal or the SAT Verbal, and the Test of Standard Written English (TSWE). Included in the report are comparative data on performance across tests and interpretive information on how combined test results might best be used in the admission process.

An Exploration of Speaking Proficiency Measures in the TOEFL Context: Clark, John L.D., and Swinton, Spencer S. Report 4. October 1979.

Describes a three-year study involving the development and experimental administration of test formats and item types aimed at measuring the English-speaking proficiency of nonnative speakers. Factor analysis and other techniques were used to identify subsets of item formats and individual items having satisfactory correlations with the Foreign Service Institute criterion interview administered to the test subjects. The results were grouped into a prototype "Test of Spoken English."

The Relationship between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language: Powers, Donald E. Report 5. December 1980.

Summarizes analyses indicating performance of 6,000 nonnative speakers of English on TOEFL and GMAT. In addition to comparisons between native and nonnative speakers, data are included showing performance by language background. A variety of analyses support the basic differences in the two tests by showing expected GMAT verbal scores for various levels of TOEFL scores.

Factor Analysis of the Test of English as a Foreign Language for Several Language Groups: Powers, Donald E., and Swinton, Spencer S. Report 6. December 1980.

Provides evidence from a set of exploratory analytical techniques that three major factors underlie performance on TOEFL. Some support is also found for concluding that these factors may be interpreted differently for several language groups. The report discusses implications for making inferences based on TOEFL subscores and considerations for future test development.

The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings: Clark, John L.D., and Swinton, Spencer S. Report 7. December 1980.

Presents the results of a study that examined the performance of foreign teaching assistants on the Test of Spoken English in relation to their classroom performance as judged by students. Also includes, for purposes of comparison, data showing performance of the same groups of teaching assistants on the Foreign Service oral interview and on TOEFL. Based on the analyses conducted in the study, TSE is shown to be a valid predictor of language abilities for nonnative English-speaking graduate teaching assistants.

Effects of Item Disclosure on TOEFL Performance: Angelis, Paul J.; Hale, Gordon A.; and Thibodeau, Lawrence A. Report 8. December 1980.

Reports the findings of a study designed to examine the effects of performance on TOEFL when a subset of items have been disclosed prior to an administration. Based on data from 16 intensive English training programs, the results indicate significant increases in performance in proportion to the number of items made available to students. Details are provided showing separate results by language group and by item type.

The above reports are currently available. Other research reports are planned. For further information about any of the TOEFL Research Reports, write to:

TOEFL Program Office
Box 899
Princeton, NJ 08541, USA